Learning mutual trust

Bikramjit Banerjee
Mathematical & Computer Sciences Department
University of Tulsa
bikram@euler.mcs.utulsa.edu

Rajatish Mukherjee
Mathematical & Computer Sciences Department
University of Tulsa
rajatish@euler.mcs.utulsa.edu

Sandip Sen
Mathematical & Computer Sciences Department
University of Tulsa
sandip@kolkata.mcs.utulsa.edu

ABSTRACT
Multiagent learning literature has looked at iterated two-player games to develop mechanisms that allow agents to learn to converge on Nash Equilibrium strategy profiles. Such equilibrium configuration implies that there is no motivation for one player to change its strategy if the other does not. Often, in general sum games, a higher payoff can be obtained by both players if one chooses not to respond optimally to the other player. By developing mutual trust, agents can avoid iterated best responses that will lead to a lesser payoff Nash Equilibrium. In this paper we consider 1-level agents (modelers) who select actions based on expected utility considering probability distributions over the actions of the opponent(s). We show that in certain situations, such stochastically-greedy agents can perform better (by developing mutually trusting behavior) than those that explicitly attempt to converge to Nash Equilibrium.

1. INTRODUCTION
The reinforcement learning techniques with performance and convergence guarantees have been developed for isolated single agents. The underlying assumption of such a proof is that the environment is stationary. Multi-agent or concurrent learning, however, violates this assumption. As a result, the standard reinforcement learning techniques (like Q-learning) are not guaranteed to converge in a multi-agent environment. The desired convergence in multiagent systems is on an equilibrium strategy-profile (collection of strategies of the agents) rather than optimal strategies for an individual agent.

The stochastic-game (or *Markov Games*) framework, a generalization of Markov Decision Processes for multiple players, has been used to model learning by agents in various domains [4, 3, 2]. In [2], two basic types of multiagent learners have been studied. The learners who do not model other agents, effectively considering them as passive parts of a non-stationary environment, are called 'independent learners' (ILs). We term these 0-level agents. In contrast to such agents, those that observe others' actions and rewards and use these explicitly in modeling them, are called 'joint-action learners' (JALs). We call these 1-level agents. Theorem 1 in [2] claims that both 0 and 1-level agents converge to equilibria in purely cooperative domains (coordination games). But their work is not extendible to general domains (general-sum games). The authors in [3] have adopted a complete-information general-sum game approach and provide a learning scheme that allows learners to converge to a mixed-strategy Nash Equilibrium in the limit.

Nash Equilibrium, however, does not guarantee that agents will obtain the best possible payoffs. Some non-Nash Equlibrium action combinations may yield better payoffs for both agents, which may be reached if the agents look ahead while selecting actions [1]. Such desirable non-myopic choices are preferred by both agents. While playing best response to other agents' current policy will lead to a deviation from such desirable solutions, restraint or mutual trust can enable players to stick to such action combinations.

In this paper we evaluate the possibility of concurrent learners converging to such desirable non-myopic action choices. While Hu and Wellman's approach is guaranteed to converge to Nash Equilibrium strategy profiles [3], independent, or even ordinary 1-level Q-learners have no such guarantees. In our previous work, we have observed that 0-level Q-learners often outperformed higher-level Q-learners in the long run even though their learning rate is slower [6]. In this paper we show that greedy modelers can, in their turn, outperform equilibrium seeking modelers in terms of the rewards received.

2. DEFINITIONS
In this section, we introduce some definitions to formulate a framework for concurrent learning.

DEFINITION 1. *A Markov Decision Process (MDP) is a quadruple* $\{S, A, T, R\}$*, where* S *is the set of states,* A *is the set of actions,* T *is the transition function,* $T : S \times A \to PD(S)$*,* PD *being a probability distribution, and* R *is the reward function,* $R : S \times A \to \mathcal{R}$.

A multiagent reinforcement-learning task can be looked upon as an extended MDP, with S specifying the joint-state of the agents, A being the joint-actions of the agents, $(A_1 \times A_2 \times \ldots A_n$ where A_i is the set of actions avaiable to the ith

agent), T as the joint state-transition function, and the reward function is redefined as $R : S \times A \rightarrow \mathcal{R}^n$. The functions T and R are usually unknown, necessitating learning. The goal of the ith agent is to find a strategy π_i that maximizes its expected sum of discounted rewards,

$$v(s, \pi_i) = \sum_{t=0}^{\infty} \gamma^t E(r_t^i | \pi_i, \pi_{-i}, s_0 = s)$$

where s_0 is the initial joint-state, r_t^i is the reward of the ith agent at time t, $\gamma \in [0, 1)$ is the discount factor, and π_{-i} is the strategy-profile of i's opponents. In [3] the ith agent learns π_{-i} simultaneously, and opts for the best response to it. Though myopically this is the best an agent can do, it may miss opportunities for receiving higher payoffs as in the well-known Prisoner's Dilemma problem [9].

Definition 2. *A bimatrix game is given by a pair of matrices, (M_1, M_2), (each of size $|A_1| \times |A_2|$) for a two-agent game, where the payoff of the ith agent for the joint action (a_1, a_2) is given by the entry $M_i(a_1, a_2)$, $\forall (a_1, a_2) \in A_1 \times A_2,\ i = 1, 2$.*

Each stage of an extended-MDP for two agents (it can be extended to n agents using n-dimensional tables instead of matrices), can be looked upon as a bimatrix game. A *zero-sum game* is a special bimatrix game where $M_1(a_1, a_2) + M_2(a_1, a_2) = 0,\ \forall (a_1, a_2) \in A_1 \times A_2$. In this paper we consider general-sum games, where the above sum is not a constant, and hence the individual payoffs of the agents for any joint-action are uncorrelated. We now define Nash equilibrium for such games.

Definition 3. *A pure-strategy Nash Equilibrium for a bimatrix game (M_1, M_2) is a pair of actions (a_1^*, a_2^*) such that*

$$M_1(a_1^*, a_2^*) \geq M_1(a_1, a_2^*) \quad \forall a_1 \in A_1$$

$$M_2(a_1^*, a_2^*) \geq M_2(a_1^*, a_2) \quad \forall a_2 \in A_2$$

In a Nash equlibrium the action chosen by each player is the best response to the opponent's current strategy and no player in this game has any incentive for unilateral deviation from its current strategy. A general-sum bimatrix game may not have any pure-strategy Nash Equilibrium.

Definition 4. *A mixed-strategy Nash Equilibrium for a bimatrix game (M_1, M_2) is a pair of probability vectors (π_1^*, π_2^*) such that*

$$\pi_1^{*\prime} M_1 \pi_2^* \geq \pi_1^{\prime} M_1 \pi_2^* \quad \forall \pi_1 \in PD(A_1)$$

$$\pi_1^{*\prime} M_2 \pi_2^* \geq \pi_1^{*\prime} M_2 \pi_2 \quad \forall \pi_2 \in PD(A_2)$$

where $PD(A_i)$ is the set of probability-distributions over the action space of the ith agent.

A significant property of mixed-strategy Nash Equilibria, is that there always exists at least one such equilibrium profile for an arbitrary finite bimatrix game [7]. Given such a bimatrix game (M_1, M_2), the mixed-strategy Nash Equilibrium, (π_1^*, π_2^*), can be computed using a quadratic programming approach as outlined in [5].

3. Q-LEARNING

A general, single-agent reinforcement learning task is an MDP, where the state transition and reward functions T and R are unknown. A simple, model-free and on-line technique for reinforcement learning is Q-learning [11]. In a stateless domain, as is the case with single-stage games studied in this paper, an independent Q-learner will have Q-values for each action a, $Q(a)$, and update them based on rewards r received from taking action a as follows:

$$Q(a) \leftarrow Q(a) + \alpha(r - Q(a))$$

where α is the learning-rate. This iteration has been proved to converge to optimal Q-values, for a particular structure of α, but independent of any particular exploration strategy provided it satisfies some general requirements. When a number of independent learners apply this algorithm, the convergence-guarantee does not hold due to the non-stationarity of the environment. However, such straightforward applications of Q-learning in multiagent systems have achieved success in the past [2, 8, 10, 12]. Our 1-level Q-learners learn Q-values, $Q(a, b)$, for each possible joint-action (a, b), using its observation of the actions of the other agents, but solely its own reward for joint-action. Thus the updation-rule used is

$$Q(a, b) \leftarrow Q(a, b) + \alpha(r - Q(a, b))$$

To allow these 1-level Q-learning agents to increasingly exploit their learned strategies, we use the Boltzmann exploration strategy, which slowly increases the exploitation probability. In this exploration scheme, the action a is selected with probability

$$\frac{e^{E(Q(a))/T}}{\sum_{a'} e^{E(Q(a'))/T}}.$$

where $E(Q(a)) = \sum_b p_b Q(a, b)$, p_b being computed as the relative-frequency measure from B's action history. Thus we call these agents "expected utility based probabilistic learners" or (EUPs). The temperature parameter T is started at a high value (causing more exploration) and then decreased over time, e.g., by muliplying with a decay factor, to increase the exploitation probability.

4. EXPERIMENTS

We experiment with 3×3 game matrices. Each agent has three actions to choose from, where a_is are the actions of agent A and b_is those of agent B. In figures 1, 2, 3 and 4 we present four such matrices. For any action combination, the top-right value in the corresponding matrix cell is the payoff to agent B and the bottom-left value is the payoff to agent A. The shaded entry in each matrix corresponds to the Nash Equilibrium strategy-profile. The action-profile that the agents prefer (greedy) and the desirable non-myopic solutions are also marked in each game-matrix.

In figure 1(left) there is a single pure Nash Equilibrium given by the action-profile $\langle a_3, b_3 \rangle$ giving a payoff of 5 to both

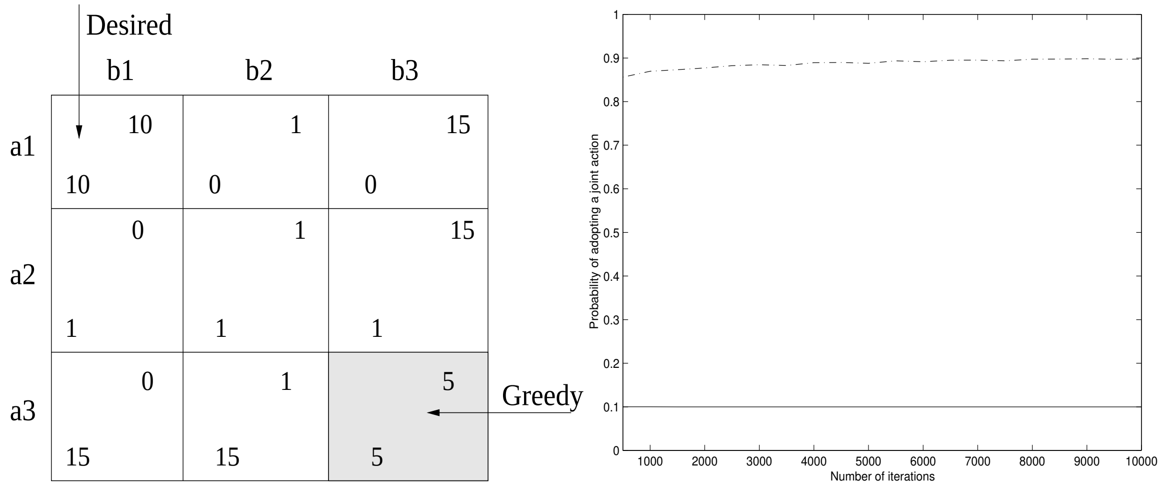

Figure 1: Game matrix where a_3 and b_3 are individually preferable to the agents, also only $\langle a_3, b_3\rangle$ is the Nash Equilibrium (left). The probability plots for the joint actions $\langle a_1, b_1\rangle$ (solid) and $\langle a_3, b_3\rangle$ are shown on the right.

agents. The desirable solution, however, is for the action-combination $\langle a_1, b_1\rangle$ giving a payoff of 10 to both agents. We used two EUPs using the above Q-learning algorithm, learning for 10,000 iterations and using 0.999942 as the temperature decay factor starting at $T = 1$. The probabilities of adopting joint-actions $\langle a_1, b_1\rangle$ and $\langle a_3, b_3\rangle$ as measured by frequencies were recorded every 500 interactions averaged over the last 500 interactions. The figures were averaged over 10 runs, and these probabilities are plotted in figure 1(right). In this case, the EUPs converge to the Nash Equilibrium in most of the runs even though the payoff is less than the desirable payoff. This is because the payoff matrix is constructed such that a_3 is the best response (actually in this example, a_3 and b_3 are also the agents' dominant strategies) of agent A irrespective of B's choice and b_3 is the best response of agent B irrespective of A's choice. However, in one run, the desirable action combination was selected by the learners.

We then reduced A's payoff for $\langle a_3, b_1\rangle$ and B's payoff for $\langle a_1, b_3\rangle$ to 9 so that both $\langle a_3, b_3\rangle$ and $\langle a_1, b_1\rangle$ are pure Nash Equilibria (figure 2(left)). However, $\langle a_1, b_1\rangle$ is the desirable solution. The corresponding probability plots are reported in figure 2(right). Here too the EUPs converge to the undesirable Nash Equilibrium and for the same reasons as listed above. The quadratic programming approach [3] produced a mixed strategy (probability distribution) of $[0, 0, 1]$ and $[0, 0, 1]$ for the agents A and B respectively. This corresponds to selecting the $\langle a_3, b_3\rangle$ action combination. Thus, our EUPs learn almost the same strategy as the mixed-strategy learners seeking Nash Equilibrium.

For the probability plot in figure 3 (right), the matrix on left has both $\langle a_1, b_1\rangle$ and $\langle a_3, b_3\rangle$ as pure Nash Equilibria. The EUPs learn to adopt the desirable action combination $\langle a_1, b_1\rangle$ in most runs. We then modified the matrix by increasing B's payoff from $\langle a_1, b_3\rangle$ to 11 (figure 4 (left)), thus leaving $\langle a_3, b_3\rangle$ as the only pure Nash Equilibrium in this matrix. From figure 4 (right) we can see that the EUPs still succeed in selecting the desirable solution more often than $\langle a_3, b_3\rangle$, even though it is not the Nash Equilibrium solution.

The profile learned by 1-level mixed strategy agent for the matrix in figure 4 (left) is $[0.09, 0, 0.91]$ and $[0.09, 0, 0.91]$ for A and B respectively. This gives an expected reward of 5.45 to each of the mixed-strategy equilibrium learners, whereas our EUPs receive expected reward of 6.3 for selection of the joint-action $\langle a_1, b_1\rangle$ alone.

The question of mutual trust can be highlighted in the matrix in figure 4 (left). If a combination of $\langle a_1, b_1\rangle$ is being played, agent B has the incentive to change its action from b_1 to b_3 to increase its payoff from 10 to 11. When it makes such a change, A's optimal response would be to change from a_1 to a_3 to increase its payoff from 4 to 5. Thus, in their haste to respond optimally to the current situation, both agents converge to an equilibrium which pays them half of what they could have got if they had showed restraint. Each of our EUPs, on the other hand, trusts the other's probability-distribution over the actions and selects its action stochastically based on that distribution. Thus they progressively tend towards the mutually beneficial part of their search space, emulating restraint which leads to mutual benefit.

5. FUTURE WORK

Our basic result is that there are certain game-structures, where stochastic modeling agents can converge to high payoff points which will be missed by sophisticated modeling learners that are designed to produce Nash Equilibrium [3]. We do not tout our empirical results as an argument for always using EUPs.

However, our observation clearly demonstrates that learning to select a Nash Equilibrium is not necessarily the best an agent can do, and that agents who are not bound by such criteria can sometimes do better. In future, we plan to study the theoretical basis for selection of a non-equilibrium solution and identify the nature and extent of mutual trust necessary to do so.

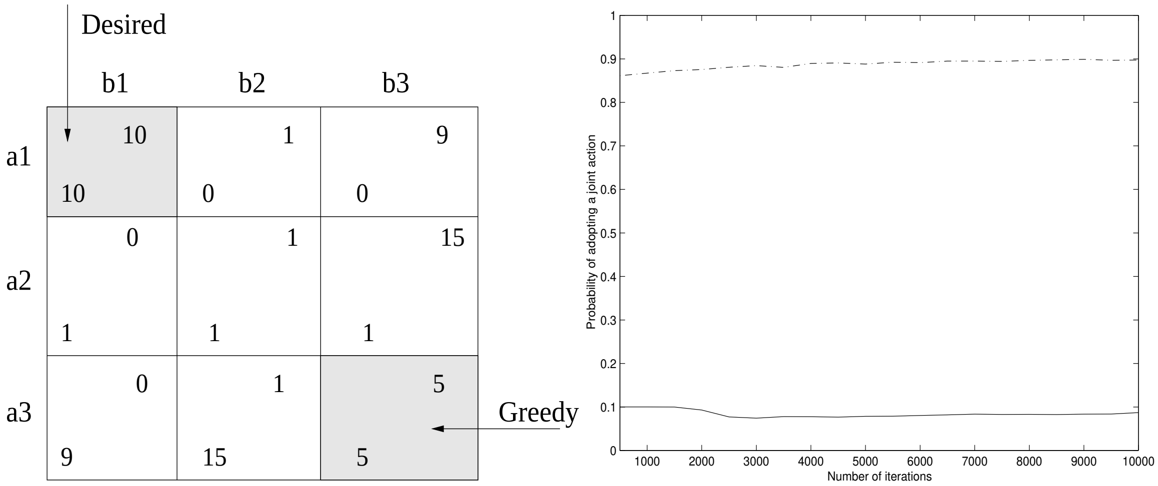

Figure 2: Game matrix where a_3 and b_3 are relatively preferable to the agents while both $\langle a_1, b_1\rangle$ and $\langle a_3, b_3\rangle$ are the Nash Equilibria (left). The probabilitry plots for the joint actions $\langle a_1, b_1\rangle$ (solid) and $\langle a_3, b_3\rangle$ are shown on the right.

We also believe that joint learners can be augmented with a greedy lookahead policy [1] rather than the best response policy (which corresponds to an immediate greedy policy) to improve their likelihood of selecting non-myopic equilibrium solutions. We plan on investigating such algorithms for discounted rewards.

Acknowledgement

This work has been supported, in part, by an NSF CAREER Award: IIS-9702672.

6. REFERENCES

[1] S. J. Brams. *Theory of moves.* Cambridge University Press, Cambridge [England] New York, USA, 1994.

[2] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, Menlo Park, CA, 1998. AAAI Press/MIT Press.

[3] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ML'98)*, pages 242–250, San Francisco, CA, 1998. Morgan Kaufmann.

[4] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, San Mateo, CA, 1994. Morgan Kaufmann.

[5] O. L. Mangasarian and H. Stone. Two-person nonzero-sum games and quadratic programming. *Journal of Mathematical Analysis and Applications*, 9:348 – 355, 1964.

[6] M. Mundhe and S. Sen. Evaluating concurrent reinforcement learners. Proceedings of the International Conference on Multiagent Systems (to appear as a poster paper), 2000.

[7] J. F. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286 – 295, 1951.

[8] T. Sandholm and R.H.Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995.

[9] T. W. Sandholm and R. H. Crites. Multiagent reinforcement learning and iterated prisoner's dilemma. *Biosystems Journal*, 37:147–166, 1995.

[10] S. Sen, M. Sekaran, and J. Hale. Learning to coordinate without sharing information. In *National Conference on Artificial Intelligence*, pages 426–431, Menlo Park, CA, 1994. AAAI Press/MIT Press. (Also published in *READINGS in AGENTS*, Michael N. Huhns and Munindar Singh (Editors), pages 509–514, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1998.).

[11] C. J. C. H. Watkins and P. D. Dayan. Q-learning. *Machine Learning*, 3:279 – 292, 1992.

[12] G. Weiß. Learning to coordinate actions in multi-agent systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 311–316, August 1993.

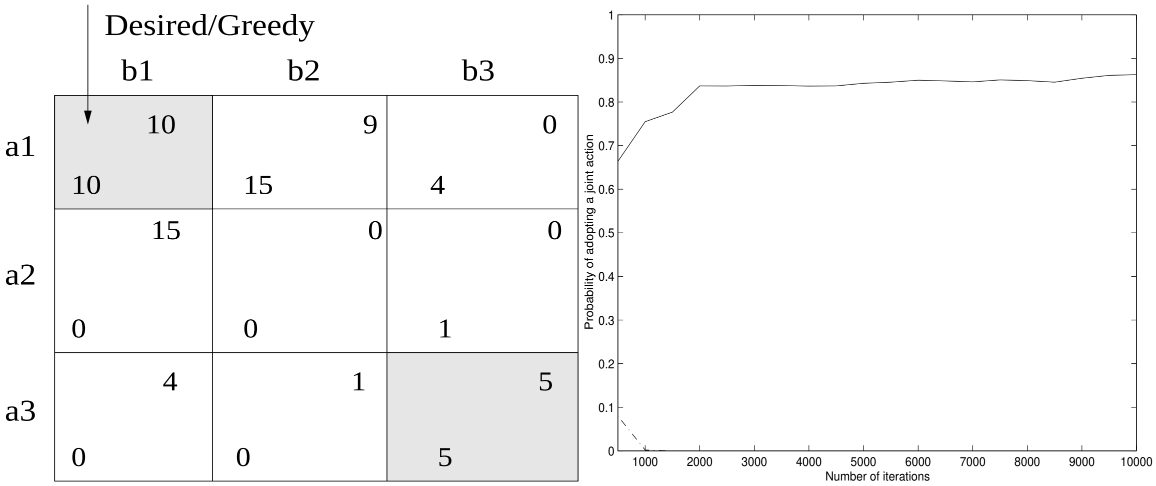

Figure 3: Game matrix where a_1 and b_1 are relatively preferable to the agents while both $\langle a_3, b_3\rangle$ and $\langle a_1, b_1\rangle$ are the Nash Equilibria (left). The probabilitry plots for the joint actions $\langle a_1, b_1\rangle$ (solid) and $\langle a_3, b_3\rangle$ are shown on the right.

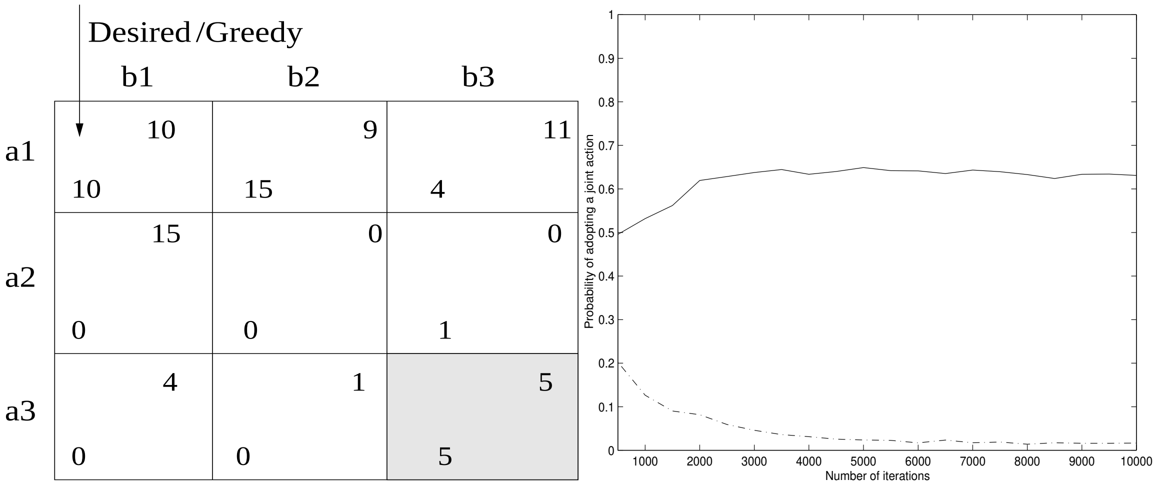

Figure 4: Game matrix where a_1 and b_1 are relatively preferable to the agents but only $\langle a_3, b_3\rangle$ is the Nash Equilibrium (left). The probabilitry plots for the joint actions $\langle a_1, b_1\rangle$ (solid) and $\langle a_3, b_3\rangle$ are shown on the right.